Orders and overlapping clusters by pyramids

Edwin Diday

▶ **To cite this version:**

Edwin Diday. Orders and overlapping clusters by pyramids. RR-0730, INRIA. 1987. inria-00075822

HAL Id: inria-00075822
https://inria.hal.science/inria-00075822

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau-
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 730

ORDERS AND OVERLAPPING CLUSTERS BY PYRAMIDS

Edwin DIDAY

OCTOBRE 1987

ORDRES ET CLASSES RECOUVRANTES PAR PYRAMIDES

ORDERS AND OVERLAPPING CLUSTERS BY PYRAMIDS

Edwin DIDAY
University of Paris IX-DAUPHINE
and
INRIA
B.P. 105
78153 LE CHESNAY CEDEX (France)

RESUME

Le problème de la recherche de classes empiétantes plutôt que de partitions se pose fréquemment dans la pratique. Les pyramides permettent une représentation visuelle de telles classes et constituent une extension naturelle des hiérarchies; elles sont plus riches que les hiérarchies du point de vue des informations fournies en faisant apparaître des recouvrements emboîtés au lieu de partitions; elles induisent un indice de dissimilarité particulier appelé "indice pyramidal" plus proche des données initiales qu'une ultramétrique. L'ensemble de ces indices (qui font intervenir la notion d'ordre sur les singletons) est en bijection avec l'ensemble des pyramides et contient l'ensemble des ultramétriques. Des procédés constructifs pour la représentation d'une pyramide sont donnés; on étudie le problème du choix optimal d'une pyramide et plus particulièrement la notion de sur-dominante et sous-dominante pyramidale. On montre comment enrichir une hiérarchie en la "pyramidisant" ou comment "hiérarchiser" partiellement une pyramide. On propose enfin d'autres formes de représentations d'un indice pyramidal (planes, polygonales, curvilignes, arbres "épais" ou "guirlandes").

ABSTRACT

The problem of overlapping clusters is frequently encountered in practice. Pyramids allow a visual representation of such clusters and constitute a natural extension of hierarchies ; they give more information than hierarchies and a more accurate order on the objects. They induce a special dissimilarity index called "pyramidal index" closer to the data than ultrametrics. The set of indices (which use the notion of order on the single objects) is in bijection with the set of pyramids and contain the set of ultrametrics. A constructive procedure for the representation of a pyramid is given ; the problem of optimal choice of a pyramid is discussed and more specifically the concept of "upper-dominant" and "sub-dominant" of pyramids. It is shown how hierarchies may be enriched by "pyramidization" and how it is possible to "hierarchize" a pyramid.

PAPIER RECUPERE ET RECYCLE

INTRODUCTION

Confronted with the multidimensional reality, man has since the beginning of time sought visual representations to understand this reality better. Hierarchical representation is a visual mode of representation which has been used through the ages ; from Aristotle and his tree of life to the recent work of Sneath and Sokhal (1973) and of Benzecri and his collaborators (1973) including Adanson (1757) and his algorithm (of hierarchical construction) to group plants, throughout history hierarchy has been a useful tool for representing multidimensional data.

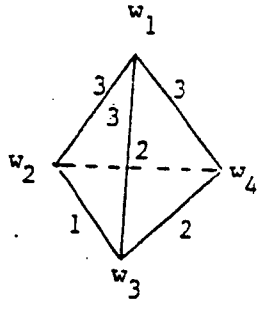
Due to the quantity and variety of practical applications, many authors have addressed themselves in recent years to research of overlapping clusters and four approaches have been mainly studied : 1) Hubert (1974) and B. Monjardet (1980) using graph theory, 2) a probability approach : a breaking down of a mixture of laws (Hartigan (1975 and 1977), Diday et al. (1979)), 3) the optimization of criteria by multidimensional scaling with Shepard and Arabie (1978) and Arabie and Caroll (1980) and by the dynamic clustering approach (Diday, Lemaire et al. (1982) : by using overlapping nuclei or weak forms), 4) a hierarchical approach with Jardine and Sibson (1971) and improvements such as those of Rohlf (1975) (in fact these are techniques leading to an improvement of the obtained hierarchies with a single link index subject to the chain effect).

Hierarchy is made up of a series of partitions. We will show in this study that all the hierarchies may be placed in a much greater scheme leading to representations (called pyramids), closer to the initial data giving a more accurate order on the objects and resulting in overlapping classes instead of partitions. The pyramid extension of hierarchies is based on the notion of compatibility between an order and a dissimilarity index ; several types of compatibility can be distinguished (see Diday in COMPSTAT) ; a natural way of

defining such a compatibility between an order Θ and a distance d is to say that three objects ordered (according to the order Θ) must be such that the distance between the extreme objects must be greater than that between the consecutive objects. It may be shown (see for instance Hubert (1974) or Diday (1983)) that if δ is an ultrametric than there exists an order compatible with δ ; the distance matrix $M(\delta,\Theta)$ which rows and columns are ordered according to order Θ is called an ultrametric matrix. The natural way of picturing this special type of matrix is to give it the shape of a hierarchy.

Example

Let $\Omega = \{w_1, w_2, w_3, w_4\}$ be a set of four objects and δ a dissimilarity index defined by figure 1 and the distance matrix $M(\delta,\theta_1)$:

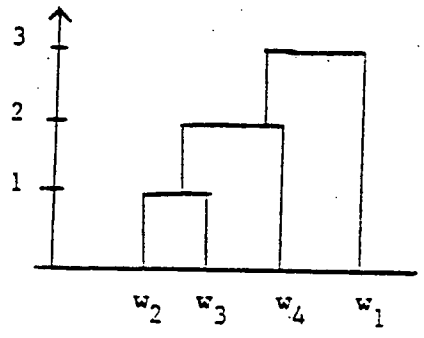


$$M(\delta,\theta_1) = \begin{array}{cccc} w_1 & w_2 & w_3 & w_4 \\ 0 & 3 & 3 & 3 \\ & 0 & 1 & 2 \\ & & 0 & 2 \\ & & & 0 \end{array} \begin{array}{l} w_1 \\ w_2 \\ w_3 \\ w_4 \end{array}$$

Figure 1

where θ_1 is the order $w_1\ w_2\ w_3\ w_4$.

θ_1 is not compatible with δ since $1 = \delta(w_2,w_3) > \delta(w_2,w_4) = 2$. The order $w_2\ w_3\ w_4\ w_1$ defines an order θ_2 which is compatible with δ ; moreover δ is an ultrametric since the ultrametric inequality $\delta(w_i,w_j) \leq \text{Max}\ \{\delta(w_i,w_k), \delta(w_k,w_j)\}\ \forall (w_i,w_j,w_k) \in \Omega^3$. Therefore the matrix $M(\delta,\theta_2)$ is ultrametric ; its hierarchical representation is given figure 2



$$M(\delta,\Theta_2) = \begin{array}{cccc|c} w_2 & w_3 & w_4 & w_1 & \\ 0 & 1 & 2 & 3 & w_2 \\ & 0 & 2 & 3 & w_3 \\ & & 0 & 3 & w_4 \\ & & & 0 & w_1 \end{array}$$

Figure 2

In this paper we extend the set of ultrametrics to a larger set which contains dissimilarities index called "pyramidal" which satisfy the compatibility condition (i.e. given a pyramidal index d there is an order Θ compatible with d) but not necessarily the ultrametric inequality. It may be shown that if Θ and d are compatible than $M(d,\Theta)$ is Robinson (i.e. rows and columns are of increasing order from the main diagonal) ; for instance, in the example given above $M(\delta,\Theta_2)$ is Robinson which is not the case of $M(\delta,\Theta_1)$. Consequently, the set of ultrametric matrices is included in the set of Robinson matrices. A natural visualisation of these Robinson matrices are pyramids which are the subject of this paper. Figure 3 which summarizes those results give an example of pyramid.

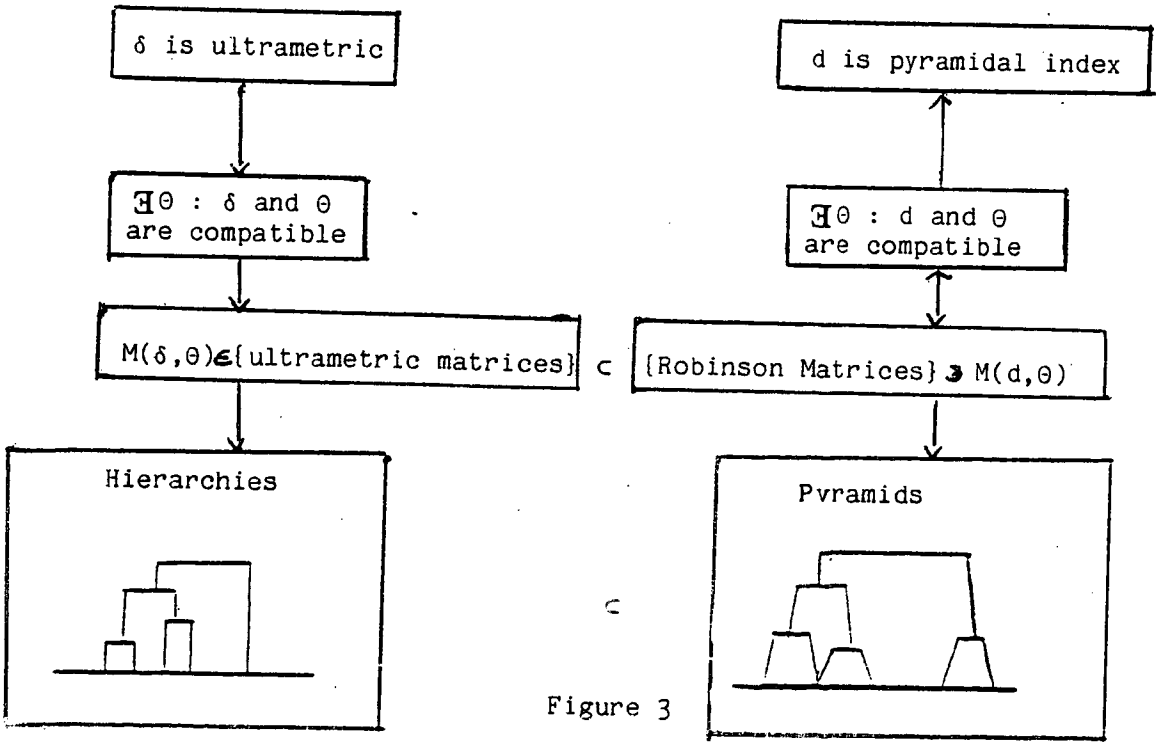


Figure 3

In paragraph 2, we give an axiomatic definition of pyramids, we show that hierarchies are specific pyramids and we give some properties concerning the visualisation of pyramids. In 3) we define the notion of indexing a pyramid which makes it possible to associate a height to each pyramid level for its visualisation. A fundamental result is given in 7) : there is a bijection between the pyramidal indices and the indexed pyramids ; this result allows us to set the problem of the search for a pyramid in optimal terms ; we deal with these problems in 5) and especially with the search for the pyramidal upper-dominant (or dominant) and sub-dominant. In contrast to ultrametrics (where the sub-dominant is an ultrametric) we show that the pyramidal sub-dominant is not necessarily a pyramidal index. In 6) we present different techniques to reduce the number of level of a pyramid ($\frac{n(n-1)}{2}$ for a saturated pyramid whereas this number is reduced to n-1 in the case of a hierarchy). Using the fact that pyramids constitute an extension of hierarchies we suggest a "hierarchization" technique which allows us to reduce the number of levels while deforming as little as possible the associated pyramidal index of a given pyramid. In the case where enriching an already constructed hierarchy is desirable to show the overlapping clusters, we suggest a pyramidization of a hierachy. Finally in 7) we give examples of output given by the computer program implementated by P. Bertrand at INRIA.

2. Definition of a pyramid and properties

2.1 Definition and link between hierarchies and pyramids

Let Ω be a finite set of objects ; we say that a part p of Ω is connex according to an order Θ, if p is an interval of this order, more precisely :
{w $\in$ p} <=> {w is between the smallest and the largest element of p according to Θ} <=> {p is connex}.
An order Θ is compatible with a set P of parts of Ω if any p $\in$ P is connex according to Θ.

Let Ω be a finite set, P a set of non-empty parts of Ω, P will be a pyramid if :

1) $\Omega \in P$ (the largest part contains all the objects),
2) $\forall w \in \Omega$, $\{w\} \in P$ (the smallest parts are the objects)
3) $\forall (p,p') \in P^2$ we have $p \cap p' = \emptyset$ or $p \cap p' \in P$
4) An order θ exists compatible with P.

Example

Let $\Omega = \{w_1,w_2,w_3\}$ and $P_1 = \{\{w_1\},\{w_2\},\{w_3\}, \{w_1,w_2\}, \{w_2,w_3\},\Omega\}$ and $P_2 = P_1 \cup \{w_1,w_3\}$. We can easily check that P_1 is a pyramid while P_2 is not, since it does not satisfy the fourth condition.

Proposition 1

The set of hierachies is included in the set of pyramids.

Proof

A hierarchy H satisfies the four conditions given in the definition of a pyramid. The two first are identical to those given in the definition of a hierarchy. For H to be a hierarchy we must have $\forall (h,h') \in H \times H$, $h \; h' = \emptyset$ or $h \cap h'$ identical to h or h', which implies that $h \cap h' = \emptyset$ or $h \cap h' \in H$. Therefore the third condition is satisfied. To prove that the fourth condition is also satisfied we can first of all construct an order on Ω, induced by H and then show that this order is compatible with H. To construct the order we use the fact that the H objects are either overlapping or of empty intersection. Starting with Ω we choose an order on the largest parts of H which are contained in Ω and we go through the procedure again with each of the parts, choosing an order on the largest part that they contain and so on, until we get the single objects which respect the induced order by this procedure and which we call θ. This order is connex for H ; to prove this, let $h \in H$ and (w',w") the extremities of h according to θ. All the elements included between w' and w" belong by construction to the parts $h_\ell \in H$ included in h and do not belong elsewhere therefore h is connex according to θ. □

2.2 Graphical representation of a pyramid

Given a pyramid P, we can say that $p \in P$ is a successor of $p' \in P$ if $p \subset p'$ (strictly) and there is not p" different from p and p', such that : $p \subset p'' \subset p'$ (strictly). We also say that p' is a predecessor of p. We know that Ω belongs to P ; the set of the successors of Ω "overlaps" Ω, since P contains the single objects. This overlapping is called a level of the pyramid. The successors of the overlapping elements contained in this level

(i.e. the successor of Ω) constitute a new level which is also an overlapping of Ω. We can go from one level to the next, and stop when the level contains only a single object since the size of the elements of P decreases, as we move down from each level to its successors.

The graphic representation of the pyramids that we have chosen, is possible because of the following properties :

1) an order Θ exists compatible with P,
2) each element of P has no more than two predecessors.

The first property satisfies the fourth condition (C4). Let us prove the second point.

Proposition 2

Each element of P has no more than two predecessors.

Proof

If p P has more than two predecessors it is equal to their intersection two by two ; otherwise it would be included in this intersection and would not be a successor. Let Θ be a compatible order with P and I the interval associated to h $\in$ P according to Θ ; suppose that h_1 , h_2 and h_3 are three predecessors of h and that I_1 , I_2 , I_3 are their associated interval according to Θ. Since h_1 and h_2 are predecessor of h and $I = I_1 \cap I_2$, I_1 and I_2 cannot contain simultaneously elements of Ω at right of I and cannot contain simultaneously elements of Ω at left of I ; otherwise, it would happen that $h_1 \subset h_2$ or $h_2 \subset h_1$ or $h_1 \cap h_2 \neq h$ and h would not be a successor of h_1 and h_2 ; if h_3 exists for the same reasons it would contain only elements at right of I or at left of I, so we would have $h_1 \subset h_3$ or $h_3 \subset h_1$ or $h_2 \subset h_3$ or $h_3 \subset h_2$ in any case h would not be simultaneously successor of h_1 , h_2 and h_3 . □

We have chosen the following graphic representation where each element of the pyramid P is represented by a horizontal segment. It is linked to its predecessors by an oblique line. Each oblique line links the middle of the segment associated with a cluster to the extremity of the segment associated with its predecessor or successor.

Figure 4

3) Pyramid indexing

Definition of an indexed pyramid

An indexed pyramid is a pair (P,f) where P is a pyramid and f an application of P in $\mathbb{R}^+$ such that :

1) f(p) = 0 if and only if p contains only one element
2) for any p and p" in P, p c p' (strict inclusion) implies f(p) ≤ f(p').

Definition of an indexed pyramid in the broad sense and strict sense

An indexed pyramid is indexed in the broad sense if :
p p' strictly and f(p) = f(p') implies the existence of p_1 and p_2 in P different from p such that p = p_1 p_2. An indexed pyramid is indexed in the strict sense if :
p c p' (strictly) implies f(p) < f(p'). The quantity f(p) is called the height of cluster p.

Définition of pyramidal index

This is a dissimilarity which satisfies the two following conditions
1) s(w,w') = 0 <=> w = w'.
2) An order θ on Ω exists such that any triplet w, w', w", with w' between w and w" according to θ satisfies the following inequality :

$$s(w, w') > \text{Max}\{s(w, w'), s(w', w")\} \text{ (pyramidal inequality)}$$

Definition of a Robinson Matrix

A matrix is Robinson if and only if the terms of the rows and columns never decrease when moving away from the main diagonal in either direction.

Proposition 3

The set of ultrametrics is included in the set of pyramidal indices.

Proof

Since an ultrametric is a distance, the first condition of the definition of a pyramidal index is satisfied by any ultrametric. We also know that as an ultrametric δ is given we can associate to it the order θ so that $M(\delta,\theta)$ is an ultrametric matrix and therefore a Robinson matrix (see for instance Diday 1983) ; the second condition is thus satisfied. □

It is also easy to show the following results.

Proposition 4

If s is a pyramidal index, then the following conditions are equivalent :

1) θ is compatible with s.
2) $M(s,\theta)$ is a Robinson matrix.
3) All paired elements (w,w') included (in the broad sense) according to θ between two elements w_i and w_j is such that $s(w_i,w_j) \geq s(w,w')$.

4. EXISTENCE OF A BIJECTION BETWEEN PYRAMIDAL INDICES AND THE INDEXED PYRAMIDS

The existence of such a bijection is an important result which extends the bijection theorem between hierarchies and ultrametrics (see for example Diday et al (1982)). The statement of this result is short but its demonstration is much longer !.

Proposition 5

There exists a bijection between the set of pyamids indexed in the broad sense π and the set of pyramidal indices S.

Proof

We will show that there exists an application ϕ of π in S and an application ψ of S in π and moreover that ϕ and ψ are inverse one from the other.

The construction of an application ϕ of π in S.

Let $\Phi : \pi \rightarrow S$ so that

$\phi((P,f) = s$ with $s(k,\ell) = \inf \{f(h)/h \in P, (k,\ell) \in h \times h\}$

where (P,f) is a pyramid noted P and indexed in the broad sense by f.

Let us show that s is a pyramidal dissimilarity.

. $s(\ell,k) = 0 \Leftrightarrow \ell = k$. This comes from the fact that $\ell = k \Rightarrow s(\ell,k) = 0$ as $s(\ell,\ell) = f(h)$ with $h = \{\ell\}$ whereby $f(h) = 0$ by definition of f ; $s(\ell,k) = 0 \Rightarrow \ell = k$ as the lowest element of P containing ℓ and k being of zero height contains only a single element of Ω by definition of f.

. $s(\ell,k) = s(k,\ell)$ as obviously $(k,\ell) \in h \times h \Rightarrow (\ell,k) \in (h \times h)$.

It remains to be demonstrated that s satisfies the pyramidal inequality or in other words, that there is an order compatible with s.

Let Θ be an order compatible with the pyramid P. We will show that Θ is compatible with s. Let us consider any triplet i,j,ℓ of Ω such that j be included between i and ℓ according to Θ. The lowest element of P noted $h_{i\ell}$ which contains i and ℓ contains i and j as it must be connex. Therefore the lowest step noted h_{ij} which contains i and j is at a maximum at the height of $h_{i\ell}$ as P is indexed in the broad sense by f, therefore $s(i,\ell) \geq s(i,j)$. By this we also show that $s(i,\ell) \geq s(j,\ell)$ whereby finally $s(i,\ell) \geq \text{Max } (s(i,j), s(j,\ell))$. Therefore Θ is compatible with s and s is a pyramidal index.

Construction of an application ψ of S in π.

Let $\psi : S \rightarrow \pi$ such that $\psi(s) = (P,f)$ where
P is the set of parts of Ω which satisfies the condition :

$$\exists \alpha : h = \{x \in \Omega \ / \ \forall \ y \in h \ s(x,y) \leq \alpha\} \qquad (1)$$

or $\{h = h_1 \ h_2$ where h_1 and h_2 are of non-empty intersection in P such that $h \neq h_1$, $h \neq h_2\}$ (2)

The set of parts h of Ω which satisfy the condition (1) is noted $P(\alpha)$.

. f is the application $P \to \mathbb{R}^+$ such that $f(h) = \text{Min } \{\alpha / h \in P(\alpha)\}$

It should be noted that $P(\alpha)$ is generally an overlapping and not a partition of Ω as the relationship $x \; R \; y \iff s(x,y) \leqq \alpha$ although reflexive and symmetrical is not transitive. It should also be noted that the set P' formed by all the elements of $P(\alpha)$ for any α in $\mathbb{R}^+$ is included and not necessarily identical to P as the intersection of elements of P' may or may not be in P' but are in P since the condition (2) must be satisfied by the elements of P ; therefore P' is not always a pyramid.

We must demonstrate that (P,f) is an indexed pyramid.

Let us first demonstrate two results which we will use later :

Lemma

1° $f(h) = \text{Min } \{\alpha / \alpha \in \mathbb{R}^+, h \in P(\alpha)\}$ in other words : $h \in P(f(h))$.
2° $f(h) = \text{Max } \{s(\ell,k) / (\ell,k) \in h \times h\}$

Proof

Let us demonstrate 1°. :

If h is reduced to one element we have $f(h) = 0$ and $h \in P(0)$ therefore $h \in P(f(h))$. If h contains more than one element, let $\alpha_0 = \inf \{\alpha / \alpha \in \mathbb{R}^+, h \in P(\alpha)\}$; Let us consider a decreasing sequence $\alpha_n \to \alpha_0$ and let ℓ an external element to h such that $s(\ell,j) = \text{Min } \{s(\kappa,i) / \kappa \notin h, i \in h, s(\kappa,i) > \alpha_0\}$, (for any $\ell \notin h$, there exists at least one $i \in h : s(\ell,i) > \alpha_0$ otherwise it would be in h). There exists a sufficiently large N so that for any $n > N$ we have $s(\ell,j) > \alpha_n$, $s(i,j) \leq \alpha_n$ for any $(i,j) \in h \times h$. In going to the limit we therefore have $s(\ell,j) > \alpha_0$ and $\forall (i,j) \in h \times h \; s(i,j) \leqq \alpha_0$. Therefore all the elements of h are in a element of $P(\alpha_0)$ and this element contains no element such ℓ which is not in h. Therefore $h \in P(\alpha_0)$ where $f(h) = \alpha_0$.

Let us demonstrate 2°. :

Let $(x,y) \in h \times h$ and $s(x,y) = \text{Max } \{s(i,j)/(i,j) \in h \times h\}$. Let $\ell \notin h$ and α_0 defined as above ; If $s(x,y) < \alpha_0$, there would exist α so that $s(x',y') < \alpha < \alpha_0 < s(\ell,y)$. We would therefore have $h \in P(\alpha)$ and α_0 would not be the smallest value so that $h \in P(\alpha_0)$. Therefore $s(x,y) \geq \alpha_0$. We have seen above in 1°) that $s(x,y) \leqq \alpha_0$ whereby finally $\alpha_0 = f(h) = s(x,y)$.

Let us now show that the pair (P,f) defines an indexed pyramid.

P satisfies the four condition which define a pyramid :

1. $\Omega \in P$. This comes from the fact that taking $\alpha = \text{Max}\{s(x,y)/(x,y) \in \Omega\times\Omega\}$. The part $h = \{x \in \Omega/ \forall y \in \Omega\ s(x,y) \leq \alpha\}$ is by definition an element of P and is identical to Ω.

2. $\forall w \in \Omega\ \{w\} \in P$. To prove this, let $h = \{w\}$, we have $h = \{w/\forall y \in h\ s(w,y) = 0\}$ as $s(w,y) = 0 \Leftrightarrow 0\ w \equiv y$ by definition of S whereby $h \in P$.

3. $h_1 \in P$, $h_2 \in P$ and $h = h_1\ h_2 \neq \emptyset$ implies $h \in P$. This comes from the fact that, if $h = h_1\ h_2$ satisfies the condition $(1)^\circ$ of the definition of P or if $h = h_1$ or $h = h_2$, h is in P and in the contrary situation $h \in P$ according to (2).

4. An order Θ compatible with P exists.

Let Θ be an order compatible with s, we will show that it is compatible with P.

Let $h \in P$, we must demonstrate that h is connex according to Θ in other words, that if w_1 and w_2 are the limits of h according to Θ we have : i) any element w included between w_1 and w_2 according to Θ is in h and ii) any element external to the interval $[w_1,w_2]$ is outside of h.

i) Let $w \in h$ such that $w_1 \leq w \leq w_2$ whereby from the proposition 4 $\forall w' \in h$ we have $s(w,w') \leq s(w_1,w_2) = \text{Max}\{s(x,y)/(x,y) \in h\times h\}$ whereby $s(w,w') \leq f(h)$. From the lemma we know that $h \in P\ (f(h))$ whereby $h = \{x \in\Omega/\forall w' \in h\ s(x,w') \leq f(h)\}$ and then $w \in h$.

ii) If $w \notin [w_1,w_2]$, w is either to the left of w_1 or to the right of w_2 according to Θ. In the first case we have : $s(w,w_2) > s(w_1,w_2) = f(h)$. In the second case : $s(w,w_1) > s(w_1,w_2) = f(h)$ therefore $w \notin h$ in any case.

Let us now demonstrate that the pyramid is indexed in the broad sense. Otherwise stated, that f satisfies the three conditions of the definition :

1° $f(h) = 0 \Leftrightarrow$ (h is a single object) ; the following proves this :

$f(h) = 0 \Leftrightarrow \min\{\alpha/h \in P\{\alpha)\} = 0$

$\Leftrightarrow \{h \in P(0)\} \Leftrightarrow \{\forall(x,y) \in h^2\ s(x,y) = 0\}$

$\Leftrightarrow \{\forall (x,y) \in h^2\ x \equiv y\} \Leftrightarrow$ {h contains only one element}

2° $(h,h') \in P\times P$ and $h \subset h'$ strictly $\Rightarrow f(h) \leq f(h')$. To prove this, let $w' \in \Omega$ contained in the complementary of h in h'. We necessarily have $s(w,w') \geq f(h')$ for at least one element $w \in h$ if not w' would be in h. In addition we know from the lemma that $s(w,w') \leq f(h') = \text{Max } \{s(\ell,k)/(\ell,k) \in h\times h\}$ as $(w,w') \in h'\times h'$ where $f(h) \leq s(w,w') \leq f(h')$. We have the inequality that we wanted.

3° $f(h) = f(h')$ and $h \subset h'$ strictly imply the existence of h_1 and h_2 in P so that $h = h_1 \cap h_2$ and $h \neq h_1$, $h \neq h_2$ because h not being identical to the largest element of P of height f(h) cannot satisfy the condition (1).

. The applications ψ and ϕ are inverse one from the other.

We need to demonstrate that $\phi \circ \psi(s) = s$ and $\psi \circ \phi((P,f)) = (P,f)$. That is demonstrate to start that $\phi \circ \psi(s) = s$, otherwise stated if $\phi((P,f)) = \sigma$ and $\psi(s) = (P,f)$ then $\sigma = s$. Let $h \in P$ of lowest height which contains x and y, by definition of σ, we know that $\sigma(x,y) = f(h)$.

Let $f(h) = \alpha$, we have $s(x,y) \leq \alpha$ as by definition of ψ, h is made up of a set of elements whose distance is less than $f(h) = \text{Min } \{\alpha'/h \in P(\alpha')\}$. As s(x,y) cannot be strictly less than α as h' containing x and y would exist less than f(h) which is contrary to the definition of h, $s(x,y) = \alpha = \sigma(x,y)$. As this result can be proven for any x and y in Ω : we have therefore $\phi o \psi(s) = s$.

Let us now demonstrate that $\psi \circ \phi((P,f)) = (P,f)$. Let $\phi((P,f)) = \sigma$ and $\psi(\sigma) = (P',f')$, we now need to demonstrate that $P \equiv P'$ and $f = f'$.

The identity between P and P' can be deduced from the following equivalency series :

(a) $\{h \in P\} \Leftrightarrow \{\{\exists (i,j) \in h\times h : h = \{x \in \Omega/\forall y \in h\ \sigma(x,y) \leq \sigma(i,j)\}$ (1) or $\{\exists h'$ and h'' in P with $h' \neq h$ and $h'' \neq h : h = h' \cap h''\}$ (2)$\}$ (b) $\Leftrightarrow \{h \in P'\}$ (c).

We will demonstrate first that (a) $\Leftrightarrow$ (b).

$\phi((P,f)) = \sigma$ and $h \in P$ imply the existence of i and j in h such that $\forall (x,y) \in h\times h$, $\sigma(x,y) \leq \sigma(i,j)$. According to proposition 6, we can choose i and j so that they constitute the extreme elements of the connex part associated with h according to an order Θ compatible with σ. As σ is pyramidal we have in fact $\sigma(i,j) \geq \sigma(x,y)$ $\forall x,y$ included between i and j according to the order Θ. If h contains all the elements w of Ω such that

$\forall y \in h\ \sigma(w,y) \leq \sigma(i,j)$ then the condition (1) is met. If not, let $w \notin h$ such that $\sigma(w,y) \leq \sigma(i,j)\ \forall y \in h$. If i is located between w and j according to Θ we have $\sigma(w,j) \geq \sigma(i,j)$ as σ is pyramidal. However by definition of w we have : $\sigma(w,j) \leq \sigma(i,j)$ whereby $\sigma(w,j) = \sigma(i,j)$ (we of course carry out the same reasoning if it were j which were located between w and i). Therefore $h' \in P$ which is of the lowest height which contains w and j is at the same height as h. As it contains w and j, it contains all the intermediate elements including those between i and j, therefore h. This results in $h \subset h'$ and $f(h) = f(h')$. By definition of a pyramid indexed in the broad sense, there therefore exists h_1 and h_2 distinct in P : $h \neq h_1$ and $h = h_1 \cap h_2$ which satisfies condition (2) ; therefore in any case (b) is satisfied.

Let us now demonstrate that (b) => (a).

We need to demonstrate that if (1) or (2) is satisfied then $h \in P$.

Let us suppose (1) to be true. Let $h' \in P$ be the lowest height which contains i and j (the two most distant points of h according to Θ) whereby $f(h') = \sigma(i,j)$ and $h \subset h'$. h' contains only elements of h as any element w external to h is such that there exists $y \in h$: $\sigma(w,y) > \sigma(i,j)$ or else it would be in h. Therefore w and y cannot simulaneously be in h' as $f(h') = \sigma(i,j)$ whereby $h' \subset h$, therefore $h' \equiv h$ which proves that h is in fact an element of P. If (2) is true h is in P by definition of a pyramid.

(b) <=> (c)

as by definition of $P(\alpha)$ we have :

$$h = \{x \in \Omega\ /\ (i,j) \in h \times h : \forall y \in h\ \sigma(x,y) \leq \sigma(i,j)\}$$

$$<=> h = \{x \in \Omega / \forall y \in h\ \sigma(x,y) \leq \alpha\} \text{ if } \alpha = \sigma(i,j)$$

(b) therefore becomes equivalent to (c) by definition of P'.

We still have to demonstrate that $f = f'$ or, that $f'(h) = f(h)$ $\forall h \in P$. For any h in $P' \equiv P$ we know from the lemma that

$$f'(h) = \text{Min}\{\alpha \in P' / h \in P(\alpha)\} => f'(h) = \sigma(i,j)$$

where $\sigma(i,j) = \text{Max}\ \{\sigma(\ell,k)/(\ell,k) \in h\times h)$. In addition, $\sigma(i,j)$ is the height of the lowest step of $P \equiv P'$ which contains i and j. As i and j are in h we therefore have $\sigma(i,j) \leqq f(h)$; let x and y be the two elements of h the furthest apart according to the order Θ. As σ is pyramidal we have $\sigma(x,y) \geqq \sigma(i,j)$ whereby $\sigma(i,j) = \sigma(x,y)$. Let us demonstrate that $\sigma(x,y) = f(h)$. If we had $\sigma(x,y) \leqq f(h)$). If we had $\sigma(x,y) \leqq f(h)$ by definition of σ, h' would be lower than h (i.e. $f(h') \leqq f(h)$) so that $\sigma(x,y) = f(h')$; but h h' as (by definition of Θ) h and h' contain all the elements included between x and y and that h contains only these. Whereby $f(h) \leqq f(h')$. It results that $f(h) = f(h') = \sigma(x,y) = \sigma(i,j)$ and finally that $\forall h \in P\ f(h) = \sigma(i,j) = f'(h)$.

□

5. CONSTRUCTING OF A PYRAMID : a pyramidal ascending classification algorithm (PAC)

By analogy with an ascending hierarchical classification and after having chosen an agregation index (this a dissimilarity index between clusters, see for example in Sneath and Sokhal (1973) a list of possible choices) we can build up a pyramidal ascending classification algorithm in the following way :

a) each element of Ω is called "group",
b) we agregate the two nearest groups among the groups which have not been agregated twice,
c) we start again with b) until a group which contains Ω is formed,
d) each time a group is formed by merging two groups we must associated an order on those two groups. Thus the algorithm builds up an order Θ on Ω,
e) two groups cannot be merged if their union is not connex,
f) let i and j be extreme elements of the connex part of Ω assiocated to a group h ; no group can be connected to a group included in h which does not contain either i or j.

Note :

The condition f) added to e) allow us to avoid crossings in the visual representation.

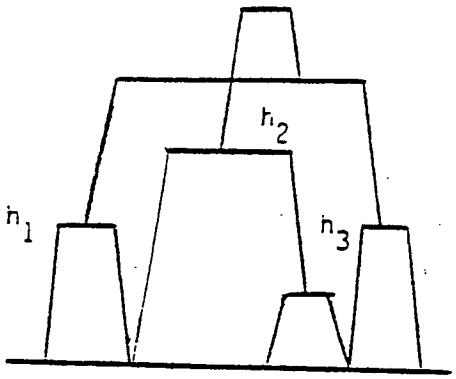


Figure 5
f ≠> e

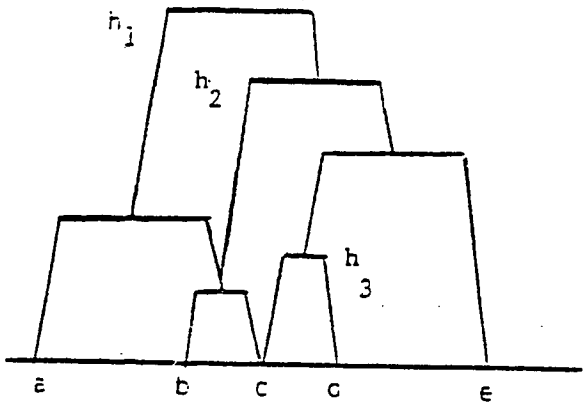


Figure 6
e ≠> f

We can see in figure 5 that the condition e) is not satisfied ($h_1 \cup h_3$ not connex due to h_2) while the condition f) is. Conversely, in figure 6, we see that condition e) is satisfied but not the condition f) as h_1 contains h_3 h_2 and h_3 do not contain either b or e the extremities of the connected part associated to h_2.

- The steps a) b) c) and the conditions d) e) f) define an algorithm called "pyramidal ascending classification" (PAC). We have the following result :

Proposition 6

The PAC algorithm builds up a pyramid.

Démonstration

Let P be the set of groups constructed by the PAC algorithm. We will show that P forms a pyramid.

To begin we can see by the construction itself that the signle objects and the set of individuals Ω are in P.

From the e) condition it is easy to show that the condition d) builds up an order Θ which is compatible with P.

We still have to show that <u>if</u> h_1 and h_2 are two groups built up by the algorithm such that $h_1 \cap h_2 \neq \emptyset$, then $h_1 \cap h_2 = h$ is in P. If h_1 and h_2 have a non-empty crossing, their common part is formed by a set of groups called $E \subset P$. Let $H_1,...H_n$ be the largest groups of E (these are groups which are not strictly contained in any group of E), we will show that these groups are in fact identical. We suppose that $h_2 \not\subset h_1$ (if $h_2 \subset h_1$ then $h_1 \cap h_2 = h_2 \in P$ thus the condition is satisfied).

Let i and j be the extremities of the order θ associated to h_1 by the algorithm (condition d)). The H_ℓ being groups of h_1 which are in h_2 must contain i or j (according to condition f).

If one of the extremities (for example j) does not appear in any of the groups, all the H_ℓ contain i. Let i' be the furthest element from i according to θ and contained in one of these groups. There therefore exists a group $H \in \{H_1,...,H_n\}$ which contains i and i'. According to the condition e) this group contains all the intermediate elements between i and i'. It contains therefore all the H_ℓ as the elements of H_ℓ are by definition of i' included between i and i' according to the order θ. As defined, the H_ℓ are not strictly included in any group E. They are all identical to H, whereby $h_1 \cap h_2 = H \in P$.

If the element i and j appear in the union of groups H_ℓ when h_2 contains (according to condition e)) all the elements included between i and j according to θ and therefore $h_1 \subset h_2$. So in this case we have also $h_1 \cap h_2 \in P$ as $h_1 \cap h_2 = h_1$. □

5. OPTIMIZATION OF A PYRAMIDAL CLASSIFICATION

6.1 Some problems of optimization

As we know that there is an equivalence between a pyramid and a pyramidal index, it is natural to raise the question of the choice of the best pyramid by searching for the pyramidal index which has the best fit with the distance d given by the user. In other words : "Optimize $\{\Delta(d,s)\ s \in S\}$" where d is the distance given by the user, S the set of pyramidal indices and Δ a measure of the fit between d and s. In the case where s is reduced to an ultrametric, different authors have studied different criteria of the following kind :

$$\Delta_1(d,s) = \sum_{i,j} p(w_i)\, p(w_j)\, (d(w_i,w_j) - s(w_i,w_j))^2$$

(Defay (1975), Caroll (1975) and Pruzansky (1975), Chandon and al (1980), Fichet (1981)).

Defays (1975) studied the following criteria

$$\Delta_2(d,s) = \sum_{i,j} |d(w_i,w_j) - s(w_i,w_j)|$$

It would be interesting to take up these studies again in the case of pyramidal indices and especially those of Fichet (1981) (which optimized the criteria Δ_1 with constraints of order).

It has been noted that the problem of finding the pyramidal index closest to d can be reduced to search for the Robinson matrix which is nearest to the dissimilarity matrix associated to d.

Hubert (1974 and 1982) suggested several criteria measuring the fit between a Robinson matrix and a dissimilarity matrix and algorithms allowing heuristic solutions to be found. In the case of few objects he purposes in these proceeding an optimal algorithm.

A classical optimization problem in hierarchical classification is that of the upper envelope of ultrametrics lower than a dissimilarity index d. It is well known that this envelope is an ultrametric (the "sub-dominant"). However the lower envelope of the ultrametrics above d (the "upper-dominant") is not necessarily an ultrametric.

It is also well known that the single link hierarchy is unique which is not always the case in the hierarchy of the complete link.

The following paragraph studies the upper and lower envelopes of d in the case of pyramidal indices.

6.2 Pyramidal sub-dominant and upper-dominant

The pyramidal sub-dominant of a dissimilarity index d is the upper-envelope of the set of pyramidal indices lower than d. We note it s_0, S being the set of pyramidal indices, and we therefore have :

$$\forall (w,w') \in \Omega^2,\ s_0(w,w') = \text{Sup}\ \{s(w,w') \leqq d(w,w'),\ s \in S\}$$
$$\text{or } s_0 = \sup\ \{s/s \leqq d,\ s \in S\}$$

We define in an analogous manner the upper-dominant noted s_u and therefore :

$$s_u = \inf\ \{s/s \geqq d,\ s \in S\}$$

A pyramid noted P_{Min} is called "single link" if it is constructed with the following agregation index :

$$\delta_{min}(h_1,h_2) = \text{Min}\ \{d(x,y)/x \in h_1,\ y \in h_2,\ x \text{ and } y \notin h_1 \cap h_2\}$$

and indexed by $f(h) = \delta_{Min}(h_1,h_2)$ where h is the ascendent of h_1 and h_2.

We will also note in the same manner P_{Max} the "complete link" pyramid build up with the following index :

$$\delta_{Max}(h_1,h_2) = \text{Max}\ \{d(x,y)/x \in h_1,\ y \in h_2\}$$

$$(\neq \text{Max}\ \{d(x,y)/x \in h_1,\ y \in h_2,\ x \text{ et } y \notin h_1\ \ h_2\})$$

We call $\phi(P,f)$ the index induced by a pyramid P indexed in the broad sense by f.

We call $s_{Min}(P)$ the index induced by a pyramid P indexed by

$f_{Min}(h) = \delta_{Min}(h_1,h_2)$. For the maximum,Max takes the place of Min.

We will demonstrate all the following results : (by using the application ϕ which has been defined by proposition 5).

- For a given pyramid P the upper envelope of the set of pyramidal indices $s = \phi(P,f)$ where f is an indexing of P which varies so that s remains less than d is s_{Min}.

In other words : if $\phi(P,f) \leqq d$ then $\phi(P,f) \leqq \phi((P,f_{Min}))$.

- We have a result which is analogous with the upper envelope :

$$\phi(P,f) \geqq \phi(P,f_{Max}) \text{ if } \phi(P,f) \geqq d.$$

- The upper envelope (respectively lower) of the set of pyramidal indices lower (respectively) then d is d which is therefore generally not a pyramidal index.

- If d is a pyramidal index then the complete link pyramid and the single link pyramid (if it is indexed in the strict sense) induce d.

All these results can be summarized in the following proposition (where remember that S is the set of pyramidal indices and d any dissimilarity index) :

Proposition 7

1) $\phi(P,f) \leq d \Rightarrow \phi(P,f) \leq \phi(P,f_{Min}) \leq d$

$\phi(P,f) \geq d \Rightarrow \phi(P,f) \geq \phi(P,f_{Max}) \geq d$

2) $\sup \{s \in S / s \leq d\} = \inf \{s \in S / s \geq d\} = d$

3) If d is a pyramidal index then $\phi(P_{Max}, f_{Max}) = d$. If in addition P_{Min} is a pyramidal index in the strict sense then $\phi(P_{Min}, f_{Min}) = d$.

Proof

1) We must prove that $\phi(P,f) \leq d \Rightarrow \phi(P,f) \leq \phi(P,f_{Min}) \leq d$.

Let us first demonstrate that $s_{Min} = \phi(P,f_{Min}) \leq d$.

Let P be a pyramid indexed by f_{Min} and $h \in P$ allowing two successors h_1 and h_2. By definition of the indexing of P we have $\forall (w_1,w_2) \in h_1 \times h_2$ with w_1 and w_2 outside of $h_1 \cap h_2$: $s(w_1,w_2) = f_{Min}(h)$. This comes from the fact that if $h' \in P$ existed containing only elements of $h'_1 = h_1 - h_1 \cap h_2$ and $h'_2 = h_2 - h_1 \cap h_2$ it would be a height superior to $f_{Min}(h) = \delta_{Min}(h_1,h_2)$ (strictly superior if h' does not contain the pair $(w'_1,w'_2) \in h'_1 \times h'_2$: $s(w'_1,w'_2) = \delta_{Min}(h_1,h_2)$, or else equal). If h' contains elements different from h'_1 and h'_2 included between w'_1 and w'_2, $f_{Min}(h')$ may be less than $f_{Min}(h)$ but then $f_{Min}(h \cup h_i) \leq f_{Min}(h)$ (with i = 1 or 2) by definition of s and we cannot have $f_{Min}(h' \cup h_i) < f_{Min}(h)$ (as then the ascendent of h_i would be $h' \cup h_i$ and not h). Therefore $\forall (w_1,w_2) \in h'_1 \times h'_2$ we have :

$$s(w_1,w_2) = f_{Min}(h) = \delta_{Min}(h_1,h_2)$$

$$= Min \{d(x,y) / x \in h_1,\ y \in h_2,\ x \text{ and } y \notin h_1 \cap h_2\} \leq d(w_1,w_2)\}$$

whereby $s = \phi(P,f_{Min}) \leq d$.

We still have to prove that $\Phi(P,f) \leq d \Rightarrow \Phi(P,f) \leq \Phi(P,f_{Min})$.

Let us take w and w' in Ω such that :

$$d(w,w') = Min \{d(x,y)/(x,y) \in h'_1 \times h'_2\} = \delta_{Min}(h_1,h_2).$$

If $s = \phi(P,f) \leq d$, we have $\forall (w'_1,w'_2) \in h'_1 \times h'_2 \quad s(w'_1,w'_2) \leq s_{Min}(w_1,w_2) = d(w,w')$ otherwise we would have $s(w,w') > d(w,w')$ (as $(w,w') \in h'_1 \times h'_2$ and s is equal to f(h) for all the pairs which are in $h'_1 \times h'_2$) which is contrary to $s \leq d$. So $s \leq s_{Min}$ and therefore $\phi(P,f) \leq \phi(P,f_{Min})$.

The implication $\phi(P,f) \geq d \Rightarrow \phi(P,f) \geq \phi(P,f_{Max}) \geq d$ can be demonstrated in a completely analogous way.

2) We must verify that

$$\sup \{s \in S/s \leq d\} = \inf \{s \in S/s \geq d\} = d.$$

Let us demonstrate first that $\sup \{s \in S/s \leq d\} = d$

Let $s^+ = \sup \{s \in S/s \leq d\}$, we must demonstrate that $\forall (w,w') \in \Omega \times \Omega \ s^+(w,w') = d(w,w')$.

Let (w,w') be any pair of $\Omega \times \Omega$; we can always build up a pyramid P with compatible order $\Theta = w_1,\ldots,w_n$ with $w_1 = w$ and $w_n = w'$ and indexed by f in the following manner : the height of the highest step $f(\Omega) = d(w,w')$, the height of the other elements of P lower than $d(w_i,w_j) = Min \{d(x,y)/(x,y) \in \Omega \times \Omega\}$. We only need to construct the lowest step h which contains w_2 and w_{n-1} at a height $f(h) \leq d(w_2,w_{n-1})$. The pyramid that is indexed in the broad sense (P,f) thus obtained induces by ϕ an index such that $s(w,w') = d(w,w')$. In addition by construction $\forall (w_\ell,w_k) \in \Omega \times \Omega$ with $(w_\ell,w_k) \neq (w_1,w_2)$ we have $s(w_\ell,w_k) \leq d(w_i,w_j) \leq d(w_\ell,w_k)$ and $s(w_1,w_2) = d(w_1,w_2)$ whereby $s \leq d$.

We therefore have $s^+(w,w') = \sup \{s(w_1,w_2)/ s \in S,\ s \leq d\} = d(w,w')$. Using the same reasoning for every pair (w,w') of Ω we have the result.

To demonstrate that $s^+ = \inf \{s \in S / s \geq d\} = d$, we use an anlogous reasoning. This time let w and w' be consecutive according to order θ. The index f is constructed in the following manner : $f(h) = d(w,w')$ where h is the lowest step of the pyramid P and contains w and w'. The other elements of P are choosen with a greater height than the greatest distance called $d(w_i,w_j)$ which separates any pair of elements of Ω. The index s induced by the pyramid indexed in the broad sense (P,f) is such that $s(w,w') = d(w,w')$ and $\forall (w_\ell,w_k) \in \Omega\times\Omega$ with $(w_\ell,w_k) \neq (w,w')$ we have $s(w_\ell,w_k) \geq d(w_i,w_j) \geq d(w_\ell,w_k)$ where finally $s \geq d$ and $s(w,w') = d(w,w')$ imply that $s^+(w,w') = d(w,w')$ for every pair of Ω.

3) We must demonstrate that $\phi(P_{Min},f_{Min}) = \phi(P_{Max},f_{Max}) = d$ if d is a pyramidal index.

It is always possible to refine an indexed pyramid (P,f) in the following way : we eliminate all the $h \in P$ such that : $f(h) = f(h')$, $h' \subset h$ and $\exists h'' \neq h : h' = h \cap h''$; the pyramid thus obtained is indexed in the broad sense. That is the case for the pyramid obtained by the CAP algorithm after refinement ; therefore $s = \phi(P_{Min},f_{Min})$ is a pyramid index if P_{Min} is refined. To demonstrate that s = d, we only need to verify that $\psi(d) = (P_{Min},f_{Min})$ as according to proposition 5 we will have $\varnothing \circ \psi(d) = d = \phi(P_{Min},f_{Min})$.

Let $(P,f) = \psi(d)$. Let h_1 and h_2 in P have the same ascendent h. We will demonstrate that $f(h) = \delta_{Max}(h_1,h_2)$. To prove this let the pair (w_i,w_j) so that $d(w_i,w_j) = \text{Max} \{d(x,y) / (x,y) \in h_1\times h_2\}$. The height of h is equal to $d(w_i,w_j)$ because it cannot be lower as we know (from the lemma of proposition 5) that $f(h) = \text{Max} \{d(x,y) / (x,y) \in h\times h\}$. It cannot be greater as there would be by definition of ψ at the level $\alpha = d(w_i,w_j)$ a connex part containing h_1 and h_2 at a lower height than that of h and therefore h would not be an ascendent of h_1 and h_2. Whereby finally

$$f(h) = d(w_i,w_j) = \text{Max} \{d(x,y)/(x,y) \in h_1\times h_2\} = \delta_{Max}(h_1,h_2)$$

We still have to prove that $f(h) = \delta_{Min}(h_1,h_2)$.

Let us suppose that the pyramid $P = \psi(d)$ is indexed in the strict sense. We have $\forall (w,w') \in h'_1\times h'_2 = (h_1-h_1\cap h_2) \times (h_2-h_1\cap h_2)$ $f(h) = d(w,w')$ if not there would be a step h' containing the element of h'_1 and h'_2 at a lower height than that of h (if $f(h') \geq f(h)$, h' would contain h by definition of $P = \psi(d)$). $h' \cap h'_1$

would give a step of the same height as h_1 and therefore the pyramid P would not be indexed in the strict sense. Or if $\forall(w,w') \in h'_1 \times h'_2$ $f(h) = d(w,w')$ we have notably

$$f(h) = \text{Min } \{d(w,w') / w \in h_1,\ w' \in h_2,\ w \text{ and } w' \notin h_1 \cap h_2\}$$

whereby $f(h) = \delta_{Min}(h_1,h_2)$.

As $f(h) = \delta_{Min}(h_1,h_2) = \delta_{Max}(h_1,h_2)$ for each ascendant h of h_1 and h_2 in P, we have $P = P_{Min} = P_{Max}$ by definition.

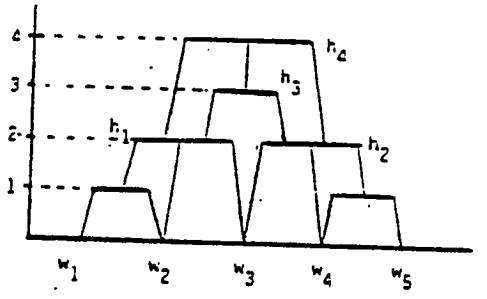
□

. $\psi(d) = (P,f)$ is certainly pyramid indexed in the strict sense if all the distances $d(w,w')$ $\forall(w,w') \in \Omega \times \Omega$ are distinct.

. We give in figure 7 an example of a pyramid index d which does not induce by ψ a pyramid of the single link as $h_4 = h_1 \cup h_2$ (whith $h_1 = \{w_1\ w_2\ w_3\}$ and $h_2 = \{w_3\ w_4\ w_5\}$) is such that $4 = f(h_4) \neq \delta_{Min}(h_1,h_2) = d(w_2,w_4) = f(h_3) = 3$.

	w_1	w_2	w_3	w_4	w_5
w_1	0	1	2	4	4
w_2		0	2	3	4
w_3			0	2	2
w_4				0	1
w_5					0

pyramidal index d



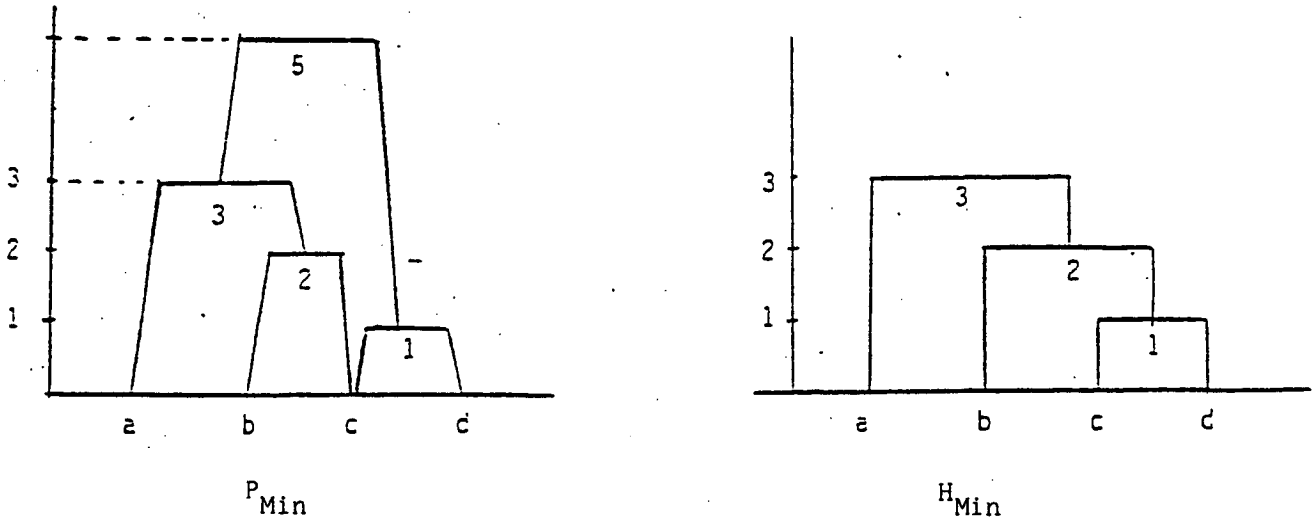
$(P,f) = \psi(d)$

Figure 7

. Let us finally note that H_{Min} (the single link hierarchy) is not necessarily included in P_{Min} (a single link pyramid). To prove this we can use the following dissimilarity index :

$$d = \begin{pmatrix} a & b & c & d & \\ 0 & 4 & 3 & 5 & a \\ & 0 & 2 & 6 & b \\ & & 0 & 1 & c \\ & & & 0 & d \end{pmatrix}$$

we obtain with d the hierarchy H_{Min} and the pyramid P_{Min} gi en in figure 8. We can then see that $h = \{b,c,d\}$ belongs to H_{Min} but not to P_{Min}.



P_{Min} H_{Min}

Figure 8

. We can also show that H_{Max} (the complete link hierarchy) is not always contained in P_{Max} (the complete link pyramid). This can be seen in taking for example the four vertices of a square with sides of lenght equal to unity.

7 HIERARCHIES AND PYRAMIDS

7.1 Saturated hierarchies and pyramids

If P is a pyramid any h ∈ P is called a "step" if h is not a single element of Ω.

Definition

A hierarchy or a pyramid is saturated when the number of its steps is maximum.

Figure 9 shows examples of saturated and non-saturated hierarchies and pyramids. From this definition we know that a hierarchy of n objects is saturated when the number of its steps is equal to n-1. A pyramid is saturated when the number of its steps is equal to the maximum number of distinct distances between objects which is $\frac{n(n-1)}{2}$. There is therefore $\frac{n}{2}$ times more steps in a saturated pyramid than in a saturated hierarchy. Therefore for n = 200 there are 100 times more steps in the saturated pyramid and the user will have difficulty in using a pyramid with $\frac{200 \times 199}{2}$ = 19,900 steps ! So as to avoid this difficulty we are lead to dealing with non-saturated pyramids.

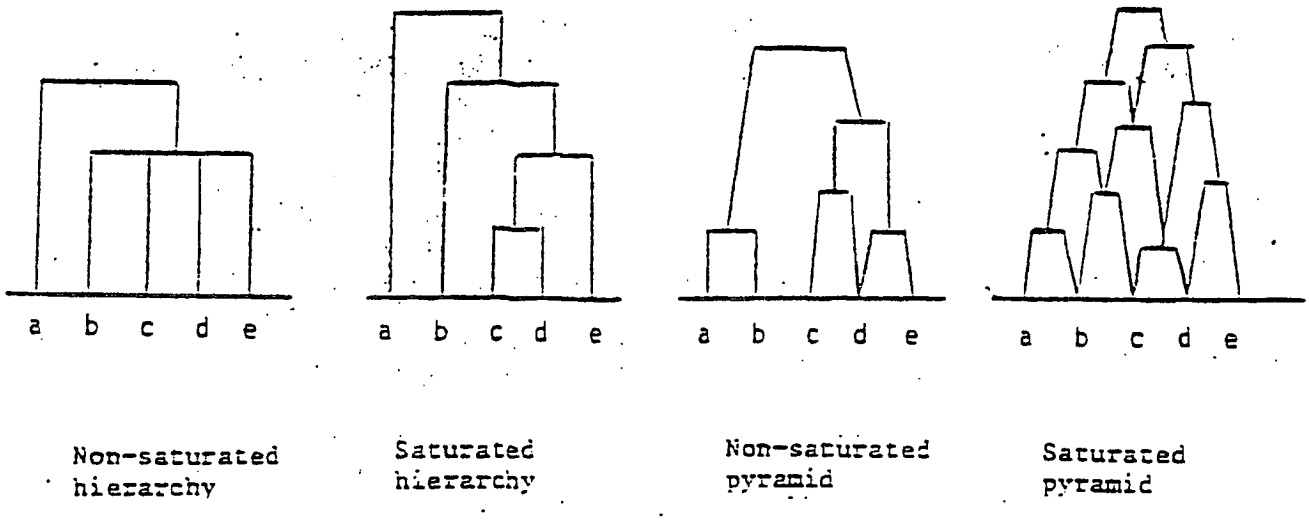


Figure 9

7.2 Construction of non-saturated pyramids

We can use two strategies : i) start with a hierarchy to enrich it with new steps which will make it pyramidal without creating "inversions" (i.e. a step must not be higher than the step in which it is contained) ; ii) start with a saturated pyramid and supress steps which are useless because they are too close. Let us see how to proceed in both cases.

i) "Pyramidization" of a hierarchy

This can be carried out in three stages : 1) we construct a hierarchy using an agregation index δ : 2) we choose an order Θ com

. Starting from the lowest we join them to form a new step each time that its height (obtained by δ) is lower than the height of the lowest step that it contains. If too many steps are obtained we use strategy ii given below, after the following example.

Example of "pyramidization"

Let us consider the dissimilarity matrix of figure 10.

	w_1	w_2	w_3	w_4
w_1	0	1	3	4
w_2		0	2	4
w_3			0	1
w_4				0

Figure 10

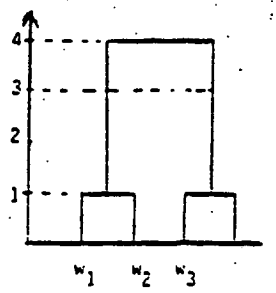


Figure 11

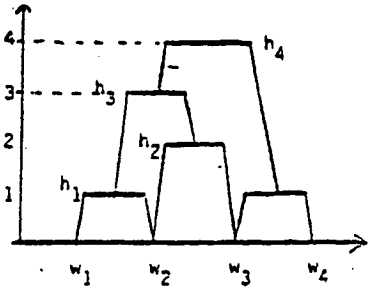


Figure 12

If we choose the maximum distance, we obtain the hierarchy of figure 11. The pyramidization of this hierarchy gives a pyramid given figure 12. The pyramid thus obtained contains two steps more than the hierarchy : the steps h_2 and h_3 which can successively appear as their height is lower than that of the lowest step that contains them : h_4.

ii) "Hierarchization" of a pyramid

A pyramid (constructed by the PAC algorithm, for example) induces a pyramidal index s and therefore a dissimilarity matrix M(s,θ) where θ is an order compatible with s. By using s we can render each triangle i j k isosceles with a smaller base than the sides by using a new dissimilarity index s' such that

$$s'(w_i,w_k) = \text{Max}\ (s(w_i,w_j),\ s(w_j,w_k))$$

if $s(w_i,w_k)$ is not the smallest side of the triangle $w_i\ w_j\ w_k$. So a pyramid becomes closer to a hierarchy each time that a triangle is made isosceles. The index s' is an ultrametric when the $C_3^n = \frac{n!}{3!(n-3)!}$ triangles are made isosceles with the base smaller than the sides.

- Algorithm of pyramid hierarchization (PH algorithm) :

1. Compare all the consecutive values on the rows and columns of the matrix M(s,θ) and retain the pair of values with the minimum difference.

2. Replace these two values by a unique value (the smallest m, the largest M, or any value included between m and M, $\frac{m+M}{2}$ for example).

We thus obtain a new dissimilarity index called s'.

3. Go back to (1) replacing s by s' as long as the percentage p of saturation of the pyramid defined by the user has not been attained.

The number of distinct values contained in the matrix $M(s,\theta)$ can be generally matched with the number of pyramid steps. (This is not true for the index of minimum proximity but is true for the maximum and most of the others).

The percentage p can thus be defined in the following manner. If there are $\frac{n(n-1)}{2}$ distinct values, the pyramid is 100 % saturated. If there are x the pyramid is saturated to $p = \frac{200\ x}{n(n-1)}$%. Consequently a hierarchy is a saturated pyramid at

$p = \frac{200(n-1)}{n(n-1)} = \frac{200}{n}$ %. The choice of p by the user allows him to define the desired "saturation rate" for the pyramid (included between 200/n and 100 if one wishes to obtain a pyramid which is not reduced to a hierarchy).

Proposition 8

At each stage of the PH algorithm the newly created index s' is a pyramidal index and the new triangle formed is isosceles with the base smaller than the sides.

Proof

Let us first note that the lenght of the sides of any triangle are the values of the matrix $M(s,\theta)$ which can be found two by two on the same row or same column (one of them will necessarily be found at the intersection of that line and that column). At the same time as $M(s,\theta)$ is a Robinson matrix the row and the columns cross at the main diagonal at the upper triangular matrix. We then know that by replacing two values : m and M consecutive (on a row or a column) by a value included between m and M, the new matrix $M(s',\theta)$ remains a Robinson matrix and therefore s' is pyramidal.

Each time that we reach a new equality in the matrix $M(s,\theta)$, we create a new isosceles triangle with a base that is maller (in the broad sense) than the sides. In fact, two consecutive values $s_{ij}\ s_{ij+1}$ on a row (resp. $s_{ij}\ s_{i+1j}$ on a column) characterize a triangle (i,j,j+1) (resp. (i,i+1,j)) whose third side (j,j+1) (resp. (i,i+1)) is of a length $s_{jj+1} \leq s_{ij+1}$ (resp. $s_{ii+1} \leq s_{ij}$) because in the superior triangular matrix $i < i+1 \leq j$ the column (resp. the lines) cross from the main diagonal. At the same time $s_{jj+1} \leq s_{ij}$ (resp. $s_{ii+1} \leq s_{i+1j}$) as s_{ij} and s_{ij+1} (resp. s_{ij} and s_{i+1j}) are the two values (consecutive and even not consecutive as $M(s,\theta)$ is a Robinson matrix) the closest by construction of the algorithm itself. Finally, the smallest side

of the triangle (i,j,j+1) (resp. (i,i+1,j) is (j,j+1) (resp. (i,i+1)). It results that equalization gives to the two largest sides of the triangle (i,j,j+1) (resp. (i,i+1,j) the same value greater than the smallest side and we thus obtain an isosceles triangle with a smaller base.

Note : Each time that we create a new equality we do not necessarily increase the number of isosceles triangles due to the equilateral triangles.

8 EXAMPLE OF PYRAMID OUTPUT

P. Bertrand has implemented a Fortran 77 program for pyramid representation at INRIA ; we give in Figure 13 an example of saturated pyramid in the case of an uniform distribution of 50 points. Figure 14 represents a low part of the saturated pyramid computed on the Ruspini data given Figure 15 and finaly in figure 16 we represent an associated non-saturated pyramid on the same data.

CONCLUSION

By looking further into the existing links between orders and ultrametrics we have formulated a theory which generalizes the theory of hierarchies and allows us to represent in visual fashion overlapping clusters. Pyramids induce orders on individuals in smaller numbers than hierarchies (2 for a saturated binary pyramid, 2^{n-1} for a binary hierarchy). To built up pyramid we can use a PAC algorithm which induces an order on Ω. We can also search for the pyramidal index s the closest to the dissimilarity index d and deduce the pyramid of the PAC with the complete link index (which in fact induces the index s). We can finally "pyramidize" a hierarchy.

The difficulty arising from the problem of too many steps in a pyramid is done away with thanks to the possibility of hierarchization and proposition 8 allows the user to specify the desired saturation rate. By "pyramidizing" a hierarchy we can also obtain a reasonable number of steps in the associated pyramid.

Many areas of research remain open : the study of pyramidal strucıures in the case where the existence of a compatible order is not imposed (in other words by supressing the fourth axiom of the definition of a pyramid) ; the study of the problem of inversions in the case of pyramids and consequences on

"pyramidal inference" should lead to more flexible and more precise agregation indices than those obtained in hierarchical inference (see Diday and Moreau 1984). It would be useful to verify that the conditions necessary to obtain a pyramid by using the accelerated algorithm of reciprocal neighbours are the same as in the hierarchical case. There is also much to be done in the search for an optimal pyramid for a given criteria (i.e. to find the pyramidal index closest to a given dissimilarity index). The links with factorial analysis need to be studied in greater depth. We can of course represent overlapping clusters (obtained by cutting a pyramid at a "significant" level) by using factorial analysis. However we can also present the following problem : the first axis of factorial analysis induces a pyramidal index, defined by the distance between the projections of individuals on this axis, it induces also a compatible order by taking account of the relative position of these projections. What is the degree of proximity between this index and the initial dissimilarity ? In which case do the pyramids give a better order ?

BIBLIOGRAPHIE

ADANSON M. (1757). "Histoire naturelle du Sénégal". Bauche ; Paris.

ARABIE P., CAROLL J.D. (1980). "MAPCLUS : a mathematical programming approach to fitting the ADCLUS model".

BENZECRI J.P., coll. (1973). "L'analyse des Données, Tome 1 La Taxonomie", Dunod.

CAROLL J.D., PRUZANSKY S. (1975). "Fitting of hierarchical tree structure", US-Japan Seminar on Multidimensional Scaling, University of California at San Diego.

CHANDON J.L., LEMAIRE J., POUGET J. (1980) "Construction de l'ultramétrique la plus proche d'une dissimilarité", RAIRO 14, 2 pp. 157-170.

DEFAYS D; (1975) "Recherche des ultramétriques à distance minimum d'une dissimilarité données". Bull de la soc. Royale des sciences de Liège, 44, 5-6, pp. 330-343.

DIDAY E. et coll. (1979). "Optimisation en classification automatique", INRIA, Rocquencourt, 78150 (France).

DIDAY E., LEMAIRE J., POUGET J., TESTU F. (1982). "Elements d'analyse des données", Dunod.

DIDAY E. MOREAU J.V. (1984) "Learning hierarchical clustering from examples" Rap. de recherche INRIA n°289.

DIDAY E. (1983). "Croisements, ordres et ultramétriques", Mathématiques des Sciences humaines ; 21ème année, n°83 pp. 31-54.

DIDAY E. (1982). "Crossing order and ultrametrics" Compstat - (Proceedings in Computer Statistics) Physica - Verlag - Vienne.

FICHET B. (1981). "Sur des approximations d'indices de dissimilarité via les représentations euclidiennes et hiérarchiques", Revue de l'ASU, statistiques et Analyse des données, Vol.2.

HARTIGAN J.A. (1975). "Clustering Algorithm", New-York : Wiley.

HARTIGAN J.A. (1977). "Clustering as modes" First International Symposimum on data analysis and Informatics, INRIA, Rocquencourt 78150 France.

HUBERT L. (1974). "Some applications on graph theory and related non-metrics techniques to problems of approximate seriation", The British Journal of Mathematical and Statistical Psychology.

HUBERT L. (1982). "Inference procedures for the evaluation and comparing of proximity matrices", Graduate School of Education UCLA.

HUBERT L. (1974). "Some applications of graph theory to clustering", Psychometrica 1974, 39, pp 283-309.

JARDINE N. and SIBSON R. (1971) "Mathematical Taxonomy". New-York : Wiley.

MONJARDET B. (1980). "Théorie des graphes et Taxonomie mathématiques", in Regards sur la théorie des graphes, presses polytechniques Romandes, Lausanne, pp. 111-125.

ROHLF F. (1975). "A new approach to the computation of the Jardine-Sibson B_k clusters", Computer Journal, 18 pp. 164-168.

SHEPARD R., ARABIE P. (1979). "Additive clustering : Representation of Similarities as Combinations of Discrete overlapping properties", Psychological Review, Vol. 86, n°2, pp. 87-123.

SNEATH P., SOKAL R. (1973). "Numerical Taxonomy", Freeman.

Figure 13

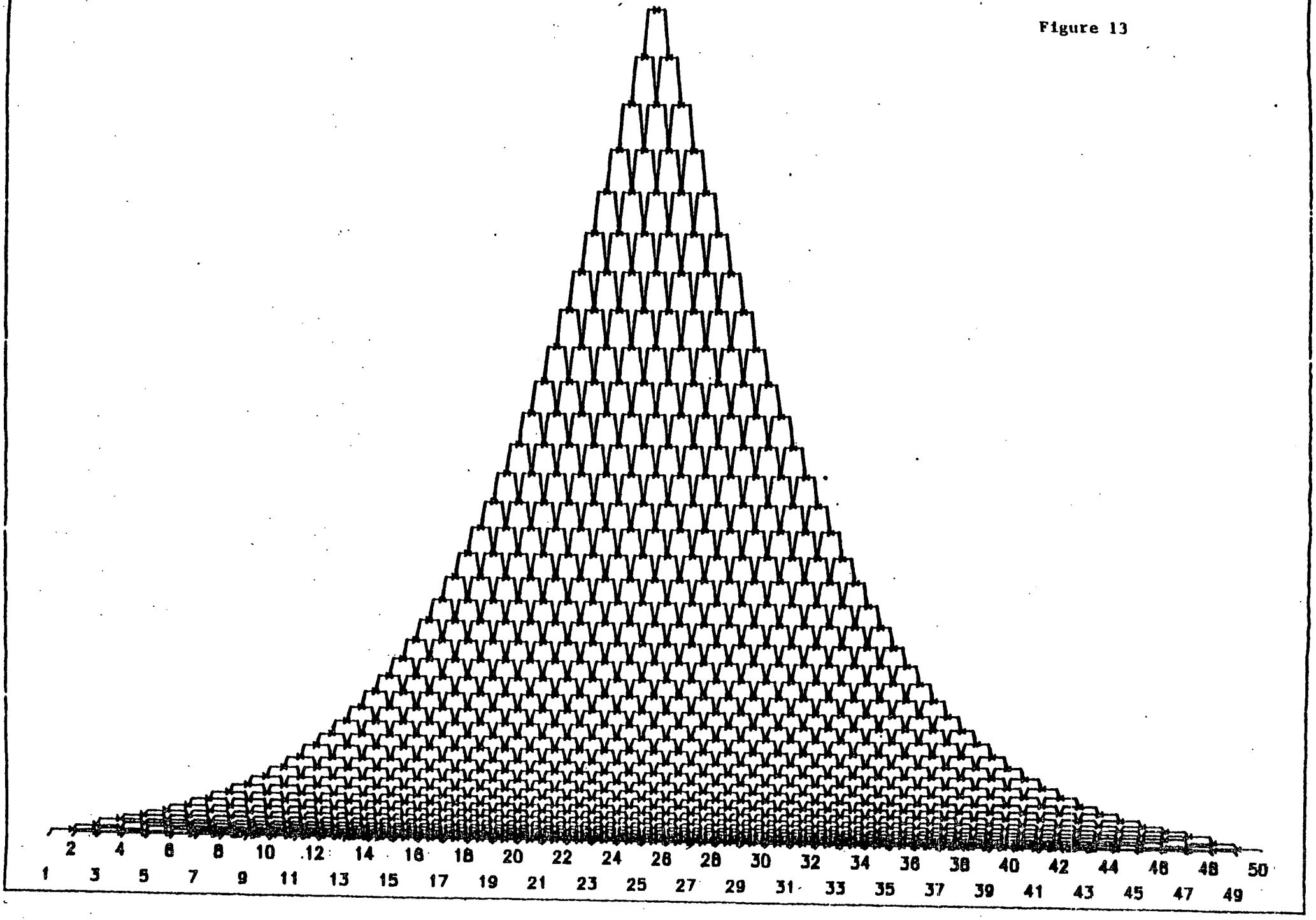
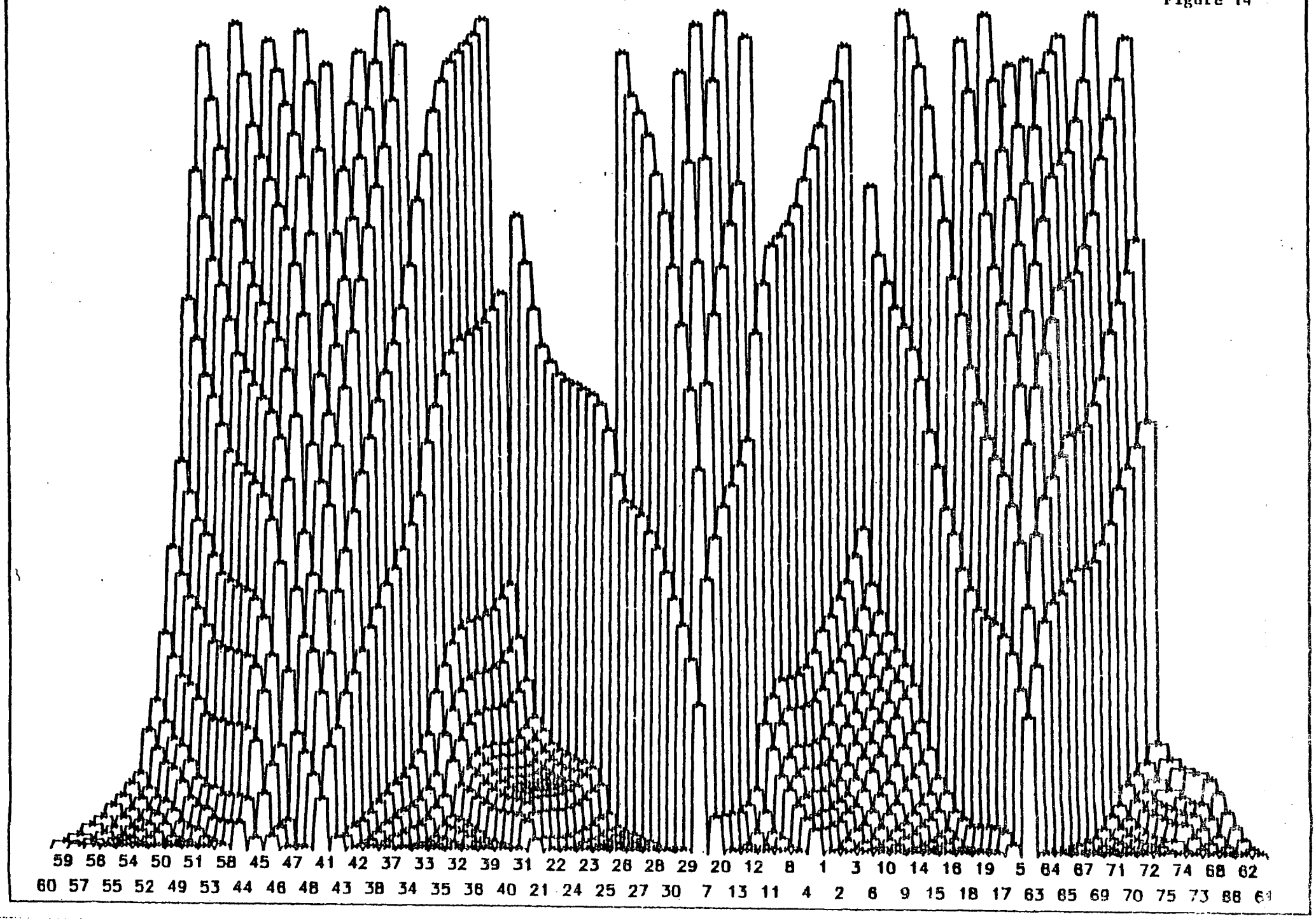

2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49

Figure 14


59 58 54 50 51 58 45 47 41 42 37 33 32 39 31 22 23 26 28 29 20 12 8 1 3 10 14 16 19 5 64 67 71 72 74 68 62
60 57 55 52 49 53 44 46 48 43 38 34 35 36 40 21 24 25 27 30 7 13 11 4 2 6 9 15 18 17 63 65 69 70 75 73 66 61

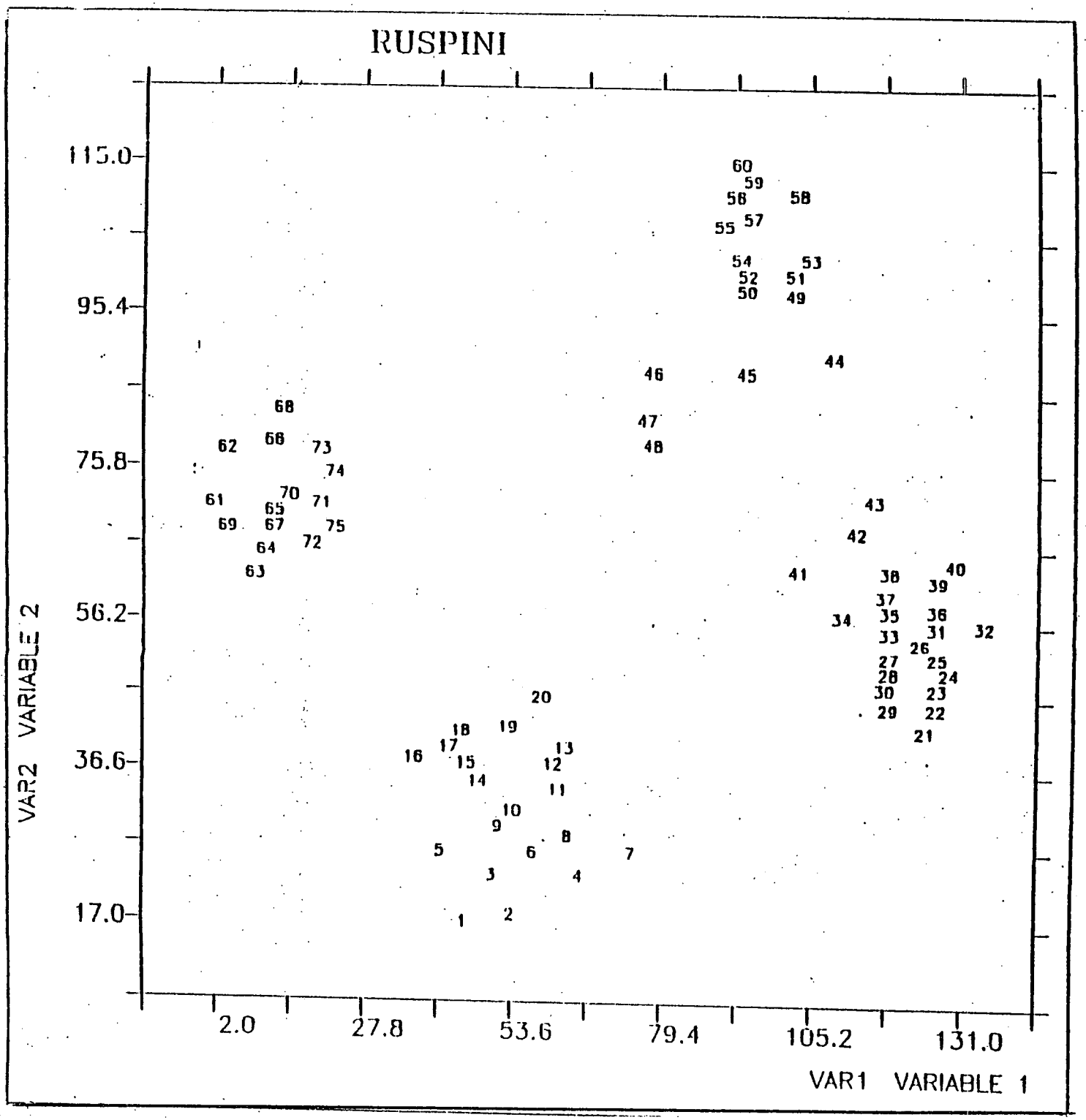
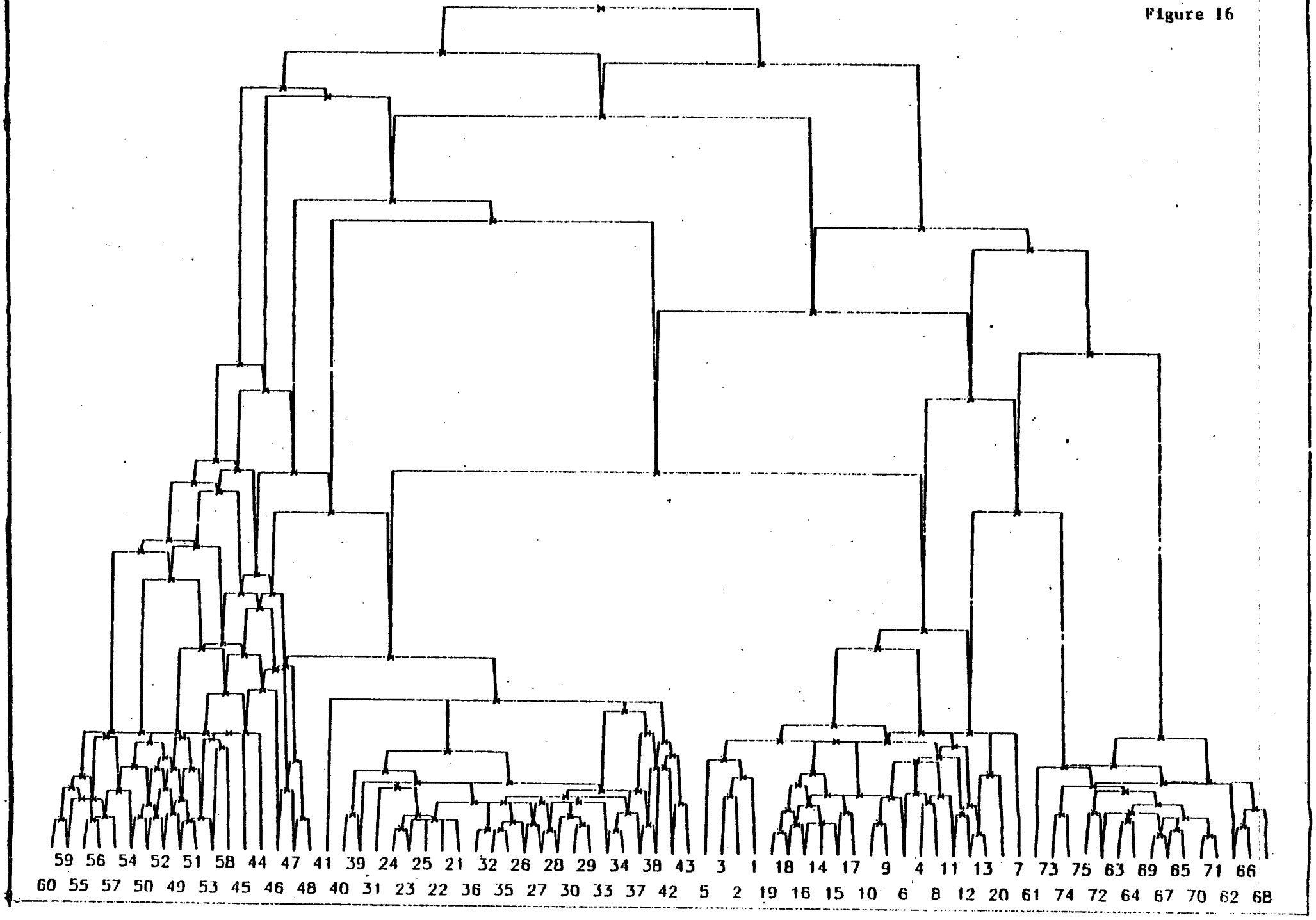

RUSPINI
115.0
95.4
75.8
56.2
36.6
17.0
VAR2 VARIABLE 2
2.0
27.8
53.6
79.4
105.2
131.0
VAR1 VARIABLE 1

Figure 15

Figure 16


59 56 54 52 51 58 44 47 41 39 24 25 21 32 26 28 29 34 38 43 3 1 18 14 17 9 4 11 13 7 73 75 63 69 65 71 66
60 55 57 50 49 53 45 46 48 40 31 23 22 36 35 27 30 33 37 42 5 2 19 16 15 10 6 8 12 20 61 74 72 64 67 70 62 68